Interpretable Classifiers for Multi-label Arrhythmia with 12-Lead Electrocardiograms

Po-Ya Hsu[1*], **Po-Han Hsu**[1], **Tsung-Han Lee**[1], **Hsin-Li Liu**[2]
[1]UC San Diego
[2]Central Taiwan University of Science and Technology

Abstract

Arrhythmia is a serious cardiovascular disease, and in recent years, several artificial intelligence programs have been proposed to automate the arrhythmia diagnosis process. However, most have not been verified on multiple datasets, and they focus on single label diagnosis. What's more concerning, these models conduct arrhythmia diagnosis in a black-box way, which prevents the cardiologists from trusting the computed results. In this study, we propose a multi-label arrhythmia classification algorithm that aims at addressing the aforementioned issues. The developed methodology is composed of three processes: selecting representation, generating features, and predicting outcomes. We developed a cache-inspired method to select a 12-lead electrocardiograms (ECG) heartbeat representation. Moreover, we devised a physiologically interpretable feature generator for segmented 12-lead ECG signals. For multi-label arrhythmia classification, we innovated an efficient arrhythmia outcome prediction procedure that is adaptable to ECG data of variant lengths. Our interpretable multi-label arrhythmia classifier was tested on six publicly available ECG datasets with over 43,000 patients' data, and our model shows the competitiveness with the ranking in the top 7% of the PhysioNet Challenge 2020.

Introduction

Arrhythmia is a serious cardiovascular disease since it has been reported to correlate with high prevalence and associated mortality (Benjamin et al. 2019). Different arrhythmia types have different mechanisms and require the appropriate interventions for successful treatments. To diagnose the arrhythmia types, cardiologists rely on the usage of electrocardiograms (ECG). ECG records the electrical activity generated from the heart, and it has been an essential tool for cardiologists to perform screening and diagnosing cardiac electrical abnormalities (Kligfield et al. 2007).

To reduce the manual arrhythmia labeling effort in ECG, several computer-aided-diagnosis (CAD) tools have been proposed. For example, an ECG waveform-based machine learning approach has been developed by Hsu and Cheng (Hsu and Cheng 2020). A deep neural network CAD has been proposed by Acharya *et al.* (Acharya et al. 2017).

*Corresponding author, email: p8hsu@eng.ucsd.edu

Although several CAD tools have been innovated, a vast amount of them has not been verified on multiple datasets. Another drawback of these CAD models is their power in diagnosis - they focus on single-label diagnosis instead of multi-label cardiac abnormalities identification. More seriously, several arrhythmia CAD algorithms perform diagnosis in a black-box manner, which could possibly discourage the cardiologists from trusting the diagnosis made by the CAD tools.

To develop an automated program that addresses the aforementioned issues, we build an interpretable multi-label arrhythmia classifier based on six ECG datasets and test our model in the PhysioNet Challenge 2020, which is a competition advocating automated, open-source approaches for classifying multi-label cardiac abnormalities from 12-lead ECGs (Perez Alday et al. 2020). We applied boosting classifier in our model to identify the cardiac abnormalities, and we deliver the computational approach that contributes to:

- Formulating 12-lead ECG heartbeat representation
- Generating physiologically reasonable feature maps
- Making efficient cardiac abnormalities identification

Problem Statement

We formulate the composite arrhythmia types diagnosis procedure into solving a multi-label classification problem. In the classification problem, the input is the 12-lead ECG data of the patient; the output is the arrhythmia types diagnosed from this patient; and our goal is to seek the medical trustworthy function that maps each input to the output.

For the sake of brevity, we introduce the following notations. We denote the input as x with the dimension of $12 \times T$, in which T is the number of the time point data in each ECG lead. For the input data sampling rate, we use f_s as the description. To describe the multi-label output, we use a binary vector y of the size $C \times 1$. Number C is equal to the total types of arrhythmia to diagnose, which is 27 in this study. We opt $0, 1$ to represent the existence of the arrhythmia types; 0 is for negative and 1 for positive. More specifically, given the fact that normal rhythm is also one of the output labels, there exists at least a 1 in every output y. We symbolize the mapping from input x to output y as h.In this study, we target at optimizing the mapping h, so that $y = h(x)$ and $L_1(y - y_{true})$ is minimized.

To quantitatively evaluate the classifier's performance, we adopt the multi-label arrhythmia scoring metric developed by the experts *et al.* (Perez Alday et al. 2020). This metric scores 100% to a perfect classifier, awards partial credit to misdiagnoses of similar outcomes or treatments, and punishes the false alarm with a negative score. According to Perez Alday *et al.*, such scoring metric can truly reflect the real clinical scenarios. The score of a classifier is expressed as

$$S = \sum_{i,j} w_{ij} a_{ij},$$

where weighting w_{ij} is developed by the cardiologists and a_{ij} represents the correctness of the classifier. More specifically, weighting matrix W is defined as

$$w_{ij} = \begin{cases} 1, & i = j, \\ \alpha, & 0 \leq \alpha < 1 \quad i \neq j, \end{cases}$$

and correctness a_{ij} is written as

$$a_{ij} = \begin{cases} \frac{1}{|\{y_p \cup c_p\}|}, & y_i = 1 \, and \quad c_j = 1, \\ 0, & otherwise, \end{cases}$$

where subscript p denotes all the positive outcomes. The final scoring metric is defined as

$$S_{metric} = \frac{S_{classifier} - S_{inactive}}{S_{true} - S_{inactive}},$$

in which the score $S_{inactive}$ is computed as a classifier that always predicts normal rhythm with all the other types negative. Our goal is to build an interpretable arrhythmia classifier that maximizes the finalized score S_{metric}.

Arrhythmia Classification Model

Our model construction includes three steps: data processing, feature generation, and model training. In data processing step, we select the datasets for model training and perform signal processing on the raw ECG signals. In feature generation, we devise a salience-based feature extraction algorithm to generate features for arrhythmia classification. In the last step, we train our model to learn the patterns of different arrhythmia types.

Data Processing

We have ECG data from six datasets across Asia, Europe, and North America, and we select four out of six to include in our model training process. The four selected datasets are G12EC, CPSC, CPSC_2, and PTB-XL databases. For every patient in each dataset, the ECG data has the dimension $12 \times T$, but some leads may have missing data with all $0s$. The dataset selection is based on the three reasons: **1) sample size**, the four datasets represent the majority of the cases; **2) data length**, variant data length, from five seconds to ten minutes, are lying within the chosen datasets; **3) signal quality**, the signal-to-noise ratio is relatively high in these datasets.

We developed a heartbeat segmentation algorithm to transform each patient's raw ECG signal into one representative heartbeat data. Such algorithm consists of four steps:

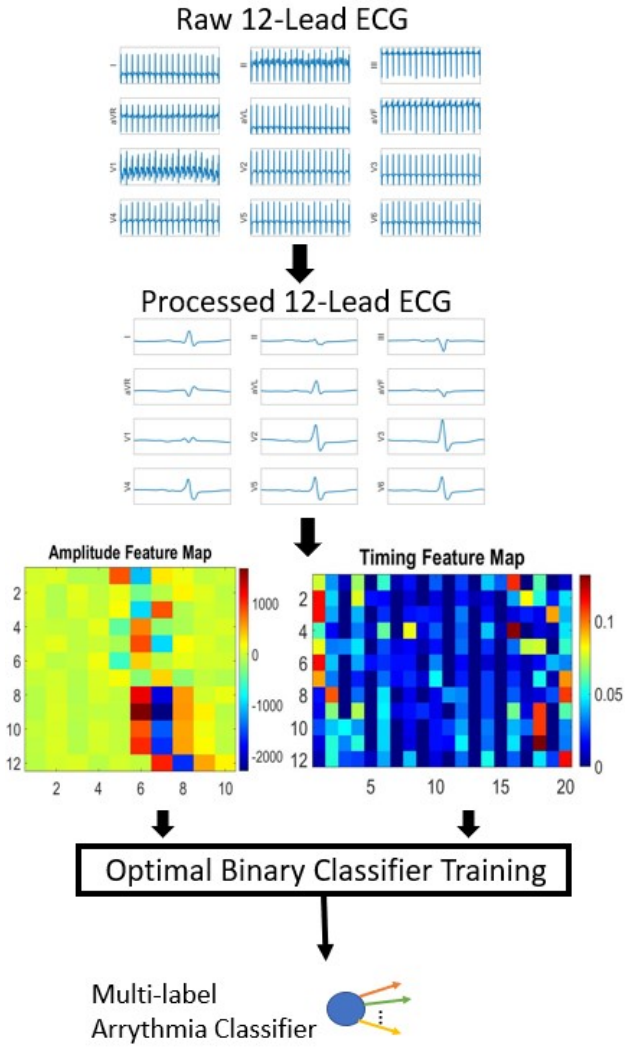

Figure 1: Flowchart of training the proposed multi-label arrhythmia classifier.

data cleaning, R-peak detection, heartbeat segmentation, and representation selection. In Figure 1, we demonstrate the raw and the processed 12-Lead ECG data.

First, we clean the raw ECG data with a Butterworth low-pass filter and a smoothing function channel-wisely. The Butterworth filter has an order of 12 and 50Hz cutoff frequency. The smoothing function adopts the moving average strategy with 10ms window.

In the second step, we detect the R-peaks in the cleaned ECG data. Most R-peaks detection is carried out on Lead II data using the famous Pan-Tompkins algorithm (Pan and Tompkins 1985). Nevertheless, for the noisy ECG data, we detect the R-peaks with the algorithm oriented for noisy physiological data as mentioned in (Chang et al. 2019).

Subsequently, for each patient, we chop down every ECG heartbeat into 1-second long frame and concatenate the frames into a tensor. Every 1-second frame has the channel-

wise R-peak located at the center. Supposed that N number of heartbeats is considered, then the tensor will have the dimension of $12 \times f_s \times N$.

Finally, we employ the clustering method and similarity metric to select the representative one-second ECG heartbeat. For each patient, we congregate the ECG frames into at most three groups and recognize the frame lying within the largest group as the representative.

The clustering approach is inspired by the cache updated rule in (Smith 1982). At the beginning, we construct a cache of three entries, and each entry contains five blocks. Next, we follow the least-recently-used rule to update our cache, which is a frequently utilized technique in computer architecture. We assign a newly visited ECG frame to an existing cache block if the ECG frame has sufficiently high similarity with the frames in the entry, or if there is an empty entry to be filled with; otherwise, we move on to the next ECG frame. Eventually, we select the block data that stores the most recent ECG frame in the largest group as the representation.

As for the similarity metric, we borrow the structural similarity index metric (SSIM) defined in (Wang et al. 2004). In this work, we empirically set the SSIM threshold as 0.3 to assign two ECG frames into the same group.

Feature Map Generation

We invented a novel physiology-inspired feature generator that is able to efficiently produce the feature maps of an arbitrary ECG frame. We incorporate the knowledge of saliency into our feature generation model to quantify the P-wave, QRS-complex, and T-wave relevant geometry on two feature maps. One feature map renders the amplitude features, while the other characterizes the timing information. We present our amplitude and timing feature generation algorithms [1,2] as follows: Both amplitude and timing generation algorithms take the 1-second 12-Lead ECG and the assumed fiducial point number K as inputs, and output the feature maps of size $12 \times K$ and $12 \times 2K$, respectively. The algorithms compute the features of each channel independently, and then project them onto the output feature maps.

Algorithm 1 Amplitude Feature Generation

Input: 1-second 12-Lead ECG Data, K
Output: Amplitude Feature Map Amp_Map

```
Initialize Amp_Map as a matrix of size 12 × K
for ch = 1 → 12 do
    x ← ECG data of channel ch
    P ← dx/dt = 0 // valleys and peaks in the data
    Q ← array of size |P| − 1
    for i = 1 → |P| − 1 do
        Q[i] = P[i + 1] − P[i]
    end for
    M ← |Q|
    I ← Indices of top K largest values in M
    // I is in non-decreasing order
    Amp_Map[ch, :] ← Q[I]
end for
```

Algorithm 2 Timing Feature Generation

Input: 1-second 12-Lead ECG Data, f_s, K
Output: Timing Feature Map $Time\_Map$

```
Initialize Time_Map as a matrix of size 12 × 2K
for ch = 1 → 12 do
    Same procedure as Algorithm 1 until obtaining I
    for i = 1 → K − 1 do
        Time_Map[ch, 2i − 1] = I[i + 1] − I[i]
        Time_Map[ch, 2i] = k − I[i]
        // I[i] < k ≤ I[i + 1], k ← argmin P − I[i]
    end for
    Time_Map[ch, 2K − 1] = f_s − I[K]
    Time_Map[ch, 2K] = k − I[K]
    // I[K] < k ≤ f_s, k ← argmin f_s − I[i]
end for
Time_Map ← Time_Map/f_s
```

For the amplitude map, the salient magnitudes are assessed; for the timing map, the durations between the salient points are taken into account. Under the assumption of existing P, Q, R, S, and T waves, we set $K = 10$ and exhibit an example in Figure 1.

Model Training

We treat solving the multi-label classification problem as training binary classifiers for each evaluated class (27 in total). Our heuristics are that each arrhythmia type bears its own unique waveform and is reflected in our generated amplitude or timing maps. Based on the reasoning, we implement the experiments detailed in the next section to train the binary classifiers for each evaluated class.

Experiments

We design three experiments to build the arrhythmia classifier that gives rise to the best performance. The three experiments are utilized to determine the ECG duration to select a heartbeat representation, the features for each arrhythmia type, and the right classification algorithm.

ECG Length: We experimented with different temporal lengths of ECG data to generate the feature maps. If the ECG data is too short, then the feature maps might not generate the ECG patterns as expected; on the other hand, if the ECG signal is lengthy, the feature maps might not capture all the representative heartbeats. Therefore, we conducted the experiments on ECG data of lengths from 5 seconds to 60 seconds, with 5 seconds as the incremental time step.

Training Features: Four experiments are carried out for feature selection: 1) amplitude feature; 2) timing feature; 3) both. 4) processed ECG data. To be more specific, we utilize not only the timing map but also the averaged heart rate and heart rate variation as our timing features.

Training Models: Basic deep learning (DL) and machine learning (ML) models are the candidates. Regarding DL strategy, convolutional neural network (CNN) and recurrent neural network (RNN) models are nominated. Referring to CNN, we borrow the AlexNet architecture with the input size being an image of 12 rows. As for RNN, we employ the

long-short term memory units with inputs having a dimension of 12. Concerning ML methods, the models examined include support vector machine, logistic regression, boosting, k-nearest neighbor, decision tree, and random forest.

To determine the best model for each evaluated class, we run the five-fold cross-validation tests on all the designed experiments. To address the underrepresented class issue, we randomly pick the samples from the class that have larger sample size to match the size of the smaller group. Furthermore, we also evaluate the performance of the classifier with the hidden test cases in the PhysioNet Challenge 2020.

Results

In this section, we briefly describe the results of each designated experiment and the performance of the final classifier.

ECG Length: We found out that 20 seconds produced the best representative ECG waveform. If the ECG length is less than 20 seconds, we often extract bad representative heartbeats in the dataset with low signal-to-noise ratio. Conversely, we neglect some representative ECG data if the chosen ECG length is larger, especially in the ECG data longer than one minute.

Based on our finding, we proposed such strategy: if the ECG data is shorter than 20s, then we construct one representative ECG following the procedure described in the model building section and determine the existence of each arrhythmia type by running through the best trained binary classifiers of each class.

For data length exceeding 20s, we randomly select N number of data segments to quantify N representations. The number N is computed by data length divided by 10 in seconds, but it is capped at 100 ($N \leq 100$). Furthermore, we determine a patient having arrhythmia type by observing $\geq 10\%$ positive labels, which was fine-tuned based on the final scoring metric.

Training Features: We discovered that the timing-deviated arrhythmia types showcase the best performance with solely the timing features. Also, we found that the abnormal waveform-based arrhythmia types favor purely the amplitude feature maps. Combining amplitude and timing feature maps does not significantly improve the performance of the classifiers. Moreover, we observed that processed ECG data as features led to the serious data over-fitting issue. The models using ECG data as inputs performed well on the specific datasets but badly on the un-trained ones.

We considered the findings of the features interesting because they support the proposed feature generation algorithm. The discovery possibly indicated that the physiologically reasonable features could describe the types of arrhythmia diagnosed.

Training Models: We sought the adaptive boosting the best multi-label arrhythmia classifier based on the robustness, generalization, and interpretability of the cross-validation tests and the hidden test cases. DL methods are inclined to over-fit the training data, and they performed terribly in the hidden cases. In addition, DL approaches favor the processed ECG data than the generated features. Other ML methods presented fairly good results with the feature maps as inputs, yet not as good as the boosting algorithm.

Scoring Metrics: We exhibit the scores of the cross-validation tests of each dataset and the hidden test cases in Table 1 as defined in *Problem Statement* section. Judging from the run-time, we show that our proposed model is competitive (compared to other competitors' models). Assessing the scores, we believe that the proposed model has successfully learned the features since the weighting scores of the cross-validation tests is similar to the official score.

Our arrhythmia classifier also showcases its competitiveness in the competition as elaborated in Table 2. In fact, we rank in the top 7% with our model that bears physiologically reasonable features computed from the ECG data.

Dataset	Runtime (hr:min:sec)	S_{metric}
CPSC	≈0:20:00	0.455
CPSC_2	≈0:30:00	0.402
G12EC	≈1:00:00	0.456
PTB	≈0:10:00	-2.589
PTB-XL	≈ 1:30:00	0.173
INCART	≈ 0:30:00	0.340
Competition Similar Data (Ours)	1:55:00	0.406
Competition Similar Data (Others')	72:00:00	≈0.3

Table 1: Performance of the proposed model.

Model	S_{metric}
Ours	**0.244**
Baseline	-0.012
Averaged Teams	0.170 [-0.65 , 0.53]

Table 2: Final scores in the competition.

Conclusion & Future Work

We deliver an interpretable multi-label arrhythmia classifier in this study. The classifier is built of our devised 12-Lead ECG heartbeat segmentation and feature generation algorithms. We demonstrate that different types of arrhythmia favor the corresponding physiological features, either the amplitude or the timing maps. We believe such features could gain more confidence from the cardiologists to trust the decisions made by our model. For future work, we aim at finding the correlations between the feature maps and the corresponding arrhythmia types.

Acknowledgments

We would like to show our gratefulness to PhysioNet Challenge 2020 organizers for generously providing the ECG datasets.

References

Acharya, U. R.; Oh, S. L.; Hagiwara, Y.; Tan, J. H.; Adam, M.; Gertych, A.; and San Tan, R. 2017. A deep convolutional neural network model to classify heartbeats. *Computers in biology and medicine* 89: 389–396.

Benjamin, E. J.; Muntner, P.; Alonso, A.; Bittencourt, M. S.; Callaway, C. W.; Carson, A. P.; Chamberlain, A. M.; Chang, A. R.; Cheng, S.; Das, S. R.; et al. 2019. Heart Disease and Stroke Statistics – 2019 Update: a Report From the American Heart Association. *Circulation* .

Chang, E.; Cheng, C.-K.; Gupta, A.; Hsu, P.-H.; Hsu, P.-Y.; Liu, H.-L.; Moffitt, A.; Ren, A.; Tsaur, I.; and Wang, S. 2019. Cuff-Less Blood Pressure Monitoring with a 3-Axis Accelerometer. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6834–6837. IEEE.

Hsu, P.-Y.; and Cheng, C.-K. 2020. Arrhythmia Classification using Deep Learning and Machine Learning with Features Extracted from Waveform-based Signal Processing. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 292–295. IEEE.

Kligfield, P.; Gettes, L. S.; Bailey, J. J.; Childers, R.; Deal, B. J.; Hancock, E. W.; Van Herpen, G.; Kors, J. A.; Macfarlane, P.; Mirvis, D. M.; et al. 2007. Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology. *Journal of the American College of Cardiology* 49(10): 1109–1127.

Pan, J.; and Tompkins, W. 1985. A real-time QRS detection algorithm. *IEEE Transaction on Biomedical Engineering* 32.

Perez Alday, E. A.; Gu, A.; Shah, A.; Robichaux, C.; Wong, A.-K. I.; Liu, C.; Liu, F.; Rad, B. A.; Elola, A.; Seyedi, S.; Li, Q.; Sharma, A.; Clifford, G. D.; and Reyna, M. A. 2020. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Under Review* .

Smith, A. J. 1982. Cache memories. *ACM Computing Surveys (CSUR)* 14(3): 473–530.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4): 600–612.